Démonstration d'un outil de « Calcul Littéraire »

Jean Rohmer 1

(1) Ecole Supérieure d'Ingénieurs Léonard de Vinci 92916 Paris La Défense Cedex
jean.rohmer@devinci.fr

Sous le nom de « Calcul Littéraire » (ou « Litteratus Calculus ») nous avons développé un démonstrateur de représentation de connaissances en langage naturel, qui peut être vu à la fois comme un gestionnaire de connaissances exprimées en langage naturel, et comme une infrastructure utile pour l'analyse de documents ou de corpus de textes. Ce travail fait suite à la réalisation d'un produit industriel à base de réseaux sémantiques, Ideiance, qui a été commercialisé par la société éponyme à partir de 1996, et utilisé dans de grandes entreprises, ainsi que par les Armées pour le Renseignement Militaire, en particulier à partir du moment où il a été repris par la société Thales en 2004. Précurseur du Web Sémantique, Idéliance permet de créer et d'exploiter collectivement, sur un serveur, des réseaux sémantiques, c'est-à-dire des énoncés de la forme Sujet / Verbe / Complément, le tout assorti de la notion de Catégorie. L'expérience a montré qu'un formalisme aussi élémentaire reste rebutant pour au moins 95% des utilisateurs, qui refusent l'effort minimal de modélisation correspondant, même habillé d'éditeurs et visualiseurs graphiques ergonomiques.

Ceci nous a conduit à imaginer un outil où les énoncés ne seraient en rien contraints, mais exprimés sous forme de phrases en langage naturel, sans autre restriction. Il sera en effet difficile à un utilisateur potentiel de dire « faire une phrase, c'est trop compliqué ou trop abstrait pour moi », comme ils le disent dès qu'il s'agit de créer un graphe sémantique et ses catégories. Le principe du « calcul littéraire » est le suivant :

On constitue un ensemble d'énoncés ou phrases indépendantes les unes des autres, appelées ***inférons***. Plus précisément un ***inféron*** est une phrase minimale et autonome compréhensible par une certaine communauté de personnes. Pour tout couple d'***inférons***, on construit leur intersection en terme de mots –après éventuelle lemmatisation-. Ces intersections s'appellent des ***interlogos***. On constitue ainsi automatiquement un graphe biparti d'***inférons*** et d'***interlogos***. A ce graphe, on peut applique un ensemble d'opérateurs visuels de navigation et contraintes ensemblistes, que nous appelons « azimuts », qui se rapprochent de la projection des graphes conceptuels, et plus généralement de la logique du premier ordre. Il faut noter que le calcul automatique des ***interlogos*** s'apparente à un mécanisme d'extraction d'entités nommées, mais sans la contrainte de disposer au préalable d'ontologies. Une fois le graphe d'***inférons*** et d'***interlogos*** constitué, on peut lui appliquer des outils de requête, de génération de tableaux et de rapports, de mise à jour habituels pour des informations structurées, aboutissant à une sorte de « ***tableur littéraire*** ».

Nous démontrerons sur une base de plus de 50 000 énoncés qu'une utilisation « brutale » du langage naturel comme format de représentation apportait un « retour sur investissement » très significatif à l'utilisateur.

Une seconde approche introduisant plus de composants linguistiques a été expérimentée et sera démontrée : elle consiste à enrichir chaque énoncé d'une analyse syntaxique et sémantique avec des analyseurs comme XIP de Xerox XRCE, ou ceux utilisés dans le projet ANR PASSAGES. On obtient des ***interlogos*** plus riches, qui permettent des navigations de phrase en phrase plus sophistiquées, reposant sur la sélection des rôles syntaxiques ou sémantiques des ***interlogos.***

Enfin, l'outil présenté peut aussi être vu comme un outil de génie linguistique, en particulier en linguistique des corpus, pour explorer un ensemble de phrases et/ou de documents, et y découvrir, par émergence, des régularités ou des différences.

Notre objectif à court terme est de développer un produit industriel, dans un cadre approprié.